Comparative visual analytics for multi-modal single-cell and spatial omics data

Basu, S.

CHAPTER 1

GENERAL INTRODUCTION

1.1 Background

Over the past decade, neuroscience has undergone a substantial transformation driven by advances in high-throughput molecular profiling technologies. In particular, single-cell RNA sequencing has enabled the systematic characterization of cellular diversity in complex tissues, including the mammalian brain. These technologies permit the measurement of genome-wide gene expression at cellular resolution, allowing the identification of transcriptionally defined cell types and states. Large-scale transcriptomic surveys have revealed extensive neuronal and non-neuronal diversity across cortical and subcortical regions.

Single-cell transcriptomic analyses of the mouse cortex have demonstrated both shared and region-specific neuronal subclasses, establishing molecular taxonomies of cortical cell types [1]. In parallel, computational integration methods have been developed to align single-cell datasets across experimental conditions, technologies, and species [2]. These advances have enabled comparative analyses that examine evolutionary conservation and divergence of cell types across organisms. Such cross-species investigations rely on orthologous gene mapping and shared low-dimensional representations to identify conserved molecular programs while highlighting species-specific innovations.

The scale of contemporary datasets has increased dramatically. Modern brain atlases may contain hundreds of thousands to millions of cells, each represented by high-dimensional gene expression vectors. These datasets frequently incorporate additional metadata, including anatomical annotations and spatial coordinates. As a consequence, the analysis and interpretation of large-scale brain data now depend critically on computational and visualization methods capable of handling high dimensionality and scale.

1.2 Visualization Challenges in Large-Scale Brain Cell Atlases

High-dimensional biological datasets require dimensionality reduction for visual exploration. Non-linear embedding techniques such as t-distributed stochastic neighbor embedding (t-SNE) [3] and Uniform Manifold Approximation and Projection (UMAP) [4, 5] are widely used to project gene expression profiles into two-dimensional spaces. These methods facilitate visual inspection of cellular heterogene-

ity and cluster structure and have become standard in single-cell analysis workflows. However, dimensionality reduction methods introduce important interpretational challenges. The transformation from high-dimensional space to a low-dimensional embedding is inherently lossy. Global distances are not always preserved, local neighborhoods may be distorted, and different parameter choices can substantially alter the resulting visualization. Consequently, visual interpretations based solely on two-dimensional embeddings must be treated with caution.

In addition to dimensionality reduction challenges, large-scale single-cell datasets present classical visualization problems associated with high-dimensional scientific data. Issues such as overplotting, perceptual clutter, scalability limitations, and cognitive overload are well documented in the visualization literature [6]. When millions of points are rendered simultaneously, meaningful structure can be obscured, and interactive exploration becomes computationally demanding.

Spatially resolved transcriptomics further increases complexity by associating molecular measurements with tissue coordinates. Spatial data often exhibit continuous gene expression gradients within discrete cell types. Capturing continuous transitions requires visualization strategies that move beyond categorical cluster representations. Moreover, when atlases from multiple species are compared, anatomical coordinate systems are not directly congruent, complicating spatial alignment and interpretation. These challenges underscore the need for visualization approaches that preserve biological interpretability while supporting scale, heterogeneity, and cross-dataset comparison.

1.3 Web-Based, Desktop-Based, and Script-Based Computational Approaches

To analyze large-scale single-cell datasets, researchers typically rely on one of three computational platforms: web-based portals, desktop applications, or script-based workflows. Each platform offers distinct capabilities suited to different research contexts.

Script-based computational workflows implemented in programming environments such as Python or R provide comprehensive pipelines for preprocessing, normalization, clustering, differential expression analysis, and visualization. Frameworks such as Scanpy [7] and Seurat [2] support reproducible workflows and enable flexible customization of statistical analyses.

Script-based environments are indispensable for method development and rigorous hypothesis testing. They allow explicit parameter control, integration of statistical models, and scalable computation on large datasets. However, visual outputs are

often generated as static figures that provide fixed views of complex data. Iterative exploration typically requires repeated code execution and regeneration of plots.

Web-based data portals provide an alternative paradigm. Such platforms enable interactive access to curated datasets through browser-based interfaces, lowering the barrier for data exploration. For large-scale brain data in particular, web-based solutions are often necessary because transferring entire datasets to local machines for analysis is impractical. Web-based platforms can also be configured for institutional or private deployment, allowing researchers to upload and analyze sensitive data within controlled environments. Several web-based approaches already support real-time exploratory interaction and visual analytics of large-scale multi-modal brain data, including transcriptomic data. However, setting up such infrastructure for specialized tasks like cross-species comparison may require more technical configuration than installing a desktop application. Maintaining large scientific data portals also requires continuous infrastructure support, database management, and long-term interface maintenance, which can impose substantial technical and operational demands on research consortia [8].

Desktop applications offer a third approach, combining local data control with interactive exploration. They provide direct access to local compute and storage resources, supporting low-latency interaction without requiring external server infrastructure. Like web-based portals and script-based workflows, desktop applications can be tailored to specific analytical tasks and implemented using modular, flexible architectures.

The choice of computational platform is primarily an engineering decision driven by project-specific requirements rather than a methodological constraint. In this context, this thesis adopts a desktop-based approach for interactive visualization and analysis while maintaining local control over data and computational resources.

1.4 Towards Interactive Visual Analytics for Comparative Analysis

Interactive visual analytics integrates automated computation with interactive visualization to support exploratory data analysis. Rather than separating computation and visualization into distinct stages, visual analytics frameworks couple algorithmic processing with dynamic user interaction. This integration allows analysts to iteratively adjust parameters, filter subsets of data, and compare alternative representations without leaving the analytical environment.

The benefits of interactive visualization have been systematically characterized in the visualization literature. Heer and Shneiderman [6] describe interaction techniques that support fluent and flexible analysis, including dynamic filtering, brushing

and linking, zooming, and coordinated multiple views. Such techniques reduce cognitive load by enabling progressive disclosure of information and supporting the rapid refinement of analytical questions.

In large-scale single-cell and comparative neuroscience datasets, interactive exploration is particularly valuable for hypothesis generation. Researchers may wish to select subsets of cells, inspect gene expression patterns across clusters, compare embeddings generated under different parameters, or examine correspondence between species. They deploy their domain knowledge to drive the exploration and interpret results on the fly, revealing novel insights in the data as the exploration progresses. Performing these tasks solely through scripted workflows can be inefficient and cognitively demanding. Interactive systems allow domain experts to interrogate uncertainty, identify outliers, and refine interpretations in real time.

The increasing scale and complexity of molecular datasets also raise broader methodological considerations. In data-rich domains such as genomics, exploratory data analysis plays a critical role in revealing unexpected structure. Yanai and Lercher [9] argue that rigid adherence to predefined hypotheses can limit discovery in complex biological systems and emphasize the value of exploratory approaches for generating new hypotheses.

In comparative brain atlas research, where millions of cells and thousands of genes are analyzed simultaneously across species, it is often not possible to anticipate all relevant patterns in advance. Therefore, interactive visual analytics may support a complementary mode of inquiry in which exploration and hypothesis formulation proceed iteratively. This approach does not replace hypothesis-driven experimentation; rather, it facilitates the generation and refinement of hypotheses grounded in large-scale observational data.

1.5 Thesis contributions and outline

This thesis develops desktop-based visual analytics approaches for interactive exploration of large-scale biological datasets, emphasizing comparative analysis across species, regions, and molecular modalities while supporting local analysis of sensitive or restricted datasets. To realize this, we investigated three specific types of comparative visualization:

Cross-species comparison forms the foundation of evolutionary neurobiology. By aligning cell types and molecular profiles across species, researchers can identify trait-specific adaptations and conserved developmental programs. This requires mapping orthologous genes, establishing consensus cell type taxonomies, and accounting for differences in brain morphology and experimental protocols across datasets.

Cross-region comparison addresses the anatomical organization of the brain. Comparing molecular profiles across cortical areas, subcortical structures, and the basal ganglia reveals regional specialization and continuous spatial gradients. However, anatomical boundaries are often gradual rather than discrete, and spatial coordinate systems vary between species, complicating direct comparisons.

Cross-modality integration connects different layers of molecular organization. Single-nucleus RNA sequencing (snRNA-seq) defines cellular identity through gene expression, while chromatin accessibility profiling (scATAC-seq) reveals regulatory elements such as promoters and enhancers. Tissue architecture is preserved in spatial transcriptomics, enabling the interpretation of laminar organization and functional domains. Integrating these modalities is essential for understanding how regulatory mechanisms drive cell type-specific expression patterns.

Methodologically, this thesis contributes task-driven visual analytics designs tailored to cross-species molecular comparison. At the system level, it introduces modular interactive components that address complementary analytical tasks, including pairwise species comparison, phylogenetic exploration, multi-modal integration, and spatial gradient analysis. These components are unified within a cohesive environment for comprehensive comparative brain atlas exploration. This thesis is situated at the intersection of computational neuroscience, data visualization, and comparative biology. It builds upon established single-cell analysis frameworks such as Scanpy [7] and Seurat [2], as well as standard dimensionality reduction techniques including t-SNE [3] and PCA [10]. Rather than replacing these approaches, the proposed methods complement them by providing an interactive layer that supports exploratory and comparative analysis at scale. As neuroscience continues to generate increasingly large and integrative atlases, effective interpretation will depend not only on statistical modeling but also on visualization paradigms that support human-centered exploration. By embedding interactive visual analytics into the analytical workflow, this thesis advances the methodological infrastructure for large-scale comparative brain research.

Throughout this thesis, the different visual analytics approaches are presented in order of increasing comparison complexity. Chapter 2 presents Cytosplore Simian Viewer, designed for comparative analysis of single-cell data of five non-human primate species for a single brain region. Chapter 3 presents Cytosplore EvoViewer, which extends single-region comparison to twenty-five species in the context of species phylogeny. Chapter 4 describes the design and validation of Cytosplore Gradient Surfer, which enables in-depth analysis and visual comparison of spatial expression gradients across multiple brain regions and species. Chapter 5 introduces multi-modal scATAC-seq integration, as well as the process of how the family of Cytosplore Viewer applications was integrated into large-scale neuroscience

discovery research studies, thereby extending the focus from method development to neuroscience applications.

Taken together, this thesis proposes the Cytosplore Viewer suite of visual analytics modules that integrates transcriptomic embeddings, spatial context, and evolutionary relationships within a unified analytical environment. It supports hierarchical navigation across anatomical structures, comparative visualization of aligned cell types, and interactive interrogation of gene expression patterns.

All software developed in the scope of this thesis is open source. The Cytosplore Viewer application and associated datasets can be downloaded from the Cytosplore Viewer website[1], and source code is available on the ManiVault Studio GitHub[2].

[1] Cytosplore Viewer Website - https://viewer.cytosplore.org
[2] Source Code Repositories - https://github.com/ManiVaultStudio